Probability Estimation over Large-Scale Random Networks via the Fiedler Delta Statistic

Antonino Freno, Mikaela Keller, Marc Tommasi

Journal of Machine Learning Research * (****) *-** Submitted */**; Published **/**

Probability Estimation over Large-Scale Random Networks via the Fiedler Delta Statistic

Antonino Freno[*] ANTONINF@AMAZON.COM
Amazon Development Center Germany GmbH
Kurfürstendamm 195
10707 Berlin, Germany

Mikaela Keller[†] MIKAELA.KELLER@INRIA.FR
Marc Tommasi[†] MARC.TOMMASI@INRIA.FR
Université Charles de Gaulle – Lille 3
Domaine Universitaire du Pont de Bois – BP 60149
59653 Villeneuve d'Ascq, France

Editor:

Abstract

Statistical models for networks have been typically committed to strong prior assumptions concerning the form of the modeled distributions. Moreover, the vast majority of currently available models are explicitly designed for capturing some specific graph properties (such as power-law degree distributions), which makes them unsuitable for application to domains where the behavior of the target quantities is not known a priori. The key contribution of this paper is twofold. First, we introduce the Fiedler delta statistic, based on the Laplacian spectrum of graphs, which allows to dispense with any parametric assumption concerning the modeled network properties. Second, we use the defined statistic to develop the Fiedler random field model, which allows for efficient estimation of edge distributions over large-scale random networks. After analyzing the dependence structure involved in Fiedler random fields, we estimate them over several real-world networks, showing that they achieve a much higher modeling accuracy than other well-known statistical approaches.

Keywords: Fiedler Delta Statistic, Fiedler Random Fields, Laplacian Spectra, Random Networks, Subgraph Sampling

1. Introduction

Arising from domains as diverse as bioinformatics and web mining, large-scale data exhibiting network structure are becoming increasingly available. Online exchange of information often tends to organize itself through some sort of network, where relevant examples include friendship networks (Facebook), customer-product networks (Amazon), co-authorship and citation networks (DBLP, Google Scholar). But although a massive application of statistical methods is a crucial element of several Web technologies such as search engines, spam filters, or recommender systems, we are still far from understanding the statistical laws

∗. This work was performed while the author was at INRIA Lille – Nord Europe.
†. Also at INRIA Lille – Nord Europe.

underlying real networks. As a consequence, a large variety of statistical models have been proposed recently, but none of them has been generally adopted as a standard reference.

Network models are commonly used to represent the relations among data units and their structural interactions. Recent studies, especially targeted at social network modeling, have focused on so-called *random graph* or *random network* models (Newman, 2010). The simplest approach is to model a random network as a configuration of binary random variables X_{uv}, such that the value of X_{uv} stands for the presence or absence of a link between nodes u and v. The general idea underlying the random graph approach is that network configurations are generated by a stochastic process governed by specific probability laws, so that different models correspond to different families of distributions over graphs (Goldenberg et al., 2009). While some of these models behave better than others in terms of computational tractability, one basic limitation affecting all of them is a sort of *parametric assumption* concerning the probability laws underlying the observed network properties. Currently available models of network structure typically assume that the shape of the probability distribution generating the network is known a priori (Erdős and Rényi, 1959; Watts and Strogatz, 1998; Albert and Barabási, 2002; Snijders et al., 2006; Leskovec et al., 2010). In such frameworks, estimating the model from data reduces to fitting the model parameters, where the parametric form of the target distribution is fixed a priori. Clearly, in order for such models to deliver accurate estimates of the distributions at hand, their prior assumptions concerning the behavior of the target quantities must be satisfied by the given data. But unfortunately, this is something that we can rarely assess a priori. To date, the knowledge we have concerning large-scale real-world networks does not allow to assess whether any particular parametric assumption is suitable for capturing the target generative process, although some observed network properties may happen to be modeled fairly well.

The aim of this paper is twofold. On the one hand, we take a first step toward non-parametric modeling of random networks by developing a novel network statistic, which we call the *Fiedler delta* statistic. The Fiedler delta function allows to model different graph properties at once in an extremely compact form. This statistic is based on the spectral analysis of the graph, and in particular on the smallest non-zero eigenvalue of the Laplacian matrix, which is known as Fiedler value (Fiedler, 1973; Mohar, 1991). On the other hand, we use the Fiedler delta statistic to define a Boltzmann distribution over graphs, leading to the *Fiedler random field* (FRF) model. Roughly speaking, for each binary edge variable X_{uv}, potentials in a FRF are functions of the difference determined in the Fiedler value by flipping the value of X_{uv}, where the spectral decomposition is restricted to a suitable subgraph incident to nodes u, v. The intuition is that the information encapsulated in the Fiedler delta for X_{uv} gives a measure of the role of X_{uv} in determining the algebraic connectivity of its neighborhood. As a first step in the theoretical analysis of FRFs, we prove that these models allow to capture edge correlations at any distance within a given neighborhood, hence defining a fairly general class of conditional independence structures over networks.

The paper is organized as follows. In Section 2 we provide a concise overview of the most widely used random network models. Section 3 reviews some theoretical background concerning the Laplacian spectrum of graphs. FRFs are then introduced in Section 4, where we also analyze their dependence structure. In Section 5 we present an efficient approach

for learning FRFs from data. To avoid unwarranted prior assumptions concerning the statistical behavior of the Fiedler delta, potentials are modeled by non-linear functions, which we estimate from data by minimizing a contrastive divergence objective. FRFs are evaluated experimentally in Section 6, showing that they are well suited for large-scale estimation problems over different network classes, while Section 7 draws some conclusions and sketches a few directions for further work.

2. Related Work

The simplest random graph model is the Erdős-Rényi (ER) model (Erdős and Rényi, 1959), which assumes that the probability of observing a link between two nodes in a given graph is constant for any pair of nodes in that graph, and it is independent of which other edges are being observed. Because of the involved independence assumption, the ER model is clearly unusable for estimating conditional distributions. Small-world models (Watts and Strogatz, 1998) try to capture such phenomena as small diameters and high clustering coefficients, which are often observed in real networks. Interestingly, the degree distribution of WS networks can be expressed in closed form in terms of two parameters δ and β, related to the average degree distribution and a network rewiring process respectively (Barrat and Weigt, 2000). In preferential attachment models (Barabási and Albert, 1999), the probability of linking to any specified node in a graph is proportional to the degree of the node in the graph, leading to "rich get richer" effects. Here, the goal is to explain the emergence of power-law degree distributions, where such distributions can be expressed in terms of an adaptive parameter α (Albert and Barabási, 2002). The Watts-Strogatz (WS) and the Barabási-Albert (BA) models have been shown to entail a Markovian and a non-Markovian dependence structure respectively (Freno et al., 2012), although such models are not explicitly designed with the goal of capturing any conditional independence structure. An explicit attempt to model potentially complex dependencies between graph edges in the form of Gibbs-Boltzmann distributions is made instead by exponential random graph (ERG) models (Snijders et al., 2006), which subsume the ER model as a special case. The two main variants of ERG approaches are the so-called Markov random graphs (MRGs) and higher-order ERGs (HRGs). These are log-linear models which differ for using as potential functions either simple triangle and (alternating) k-star counts, or the slightly more complex (alternating) k-triangle and k-star counts, respectively (Snijders et al., 2006). Based on the chosen potential functions, MRGs and HRGs are able to model edge correlations for pairs of nodes which are either contiguous or separated by at most one edge respectively (Robins et al., 2007). The parameters of all such models can be estimated by standard gradient descent (using standard maximum-likelihood or pseudo-likelihood approaches), and they can then be used to predict conditional edge distributions, exploiting either the respective potential functions or information from the degrees observed in the given subgraphs in the case of WS and BA models (Newman, 2001; Barabási et al., 2002; Freno et al., 2012).

A different approach to network representation is taken instead by (mixed-membership) stochastic blockmodels (Airoldi et al., 2008). These posit a latent set of clusters for the observed network, so that the linking behavior of the nodes is determined by their membership into one or more clusters. While the idea of a hidden structure explaining the observed links is shared by other latent space approaches (Hoff et al., 2002), one distinguishing feature of

mixed-membership stochastic blockmodels is given by the use of variational approximations in order to achieve scalability. As compared to the other approaches, stochastic blockmodels are less severely constrained by ad hoc empirical hypotheses. On the other hand, the scale of problems they have been applied to so far is still limited to networks with a few hundreds of nodes, and it is not known whether the involved tractability/accuracy tradeoffs would be still tolerable on a more realistic scale (Goldenberg et al., 2009). Finally, another attempt at modeling real networks through a stochastic generative process is made by stochastic Kronecker graphs (SKGs), which try to capture phenomena such as heavy-tailed degree distributions and shrinking diameter properties while paying attention to the temporal dynamics of network growth (Leskovec et al., 2010). One limitation of SKGs from the point of view of probabilistic inference is that no expression has been derived from them for conditional distributions of edges, conditioning e.g. on their neighborhoods. This prevents from using them for example in (conditional) link prediction applications, whereas they have been mostly used for graph generation.

3. Graphs, Laplacians, and Eigenvalues

Let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ be an undirected graph with n nodes. In the following we assume that the graph is unweighted with adjacency matrix $\mathbf{A}$. The degree d_u of a node $u \in \mathcal{V}$ is defined as the number of connections of u to other nodes, that is $d_u = |\{v : \{u, v\} \in \mathcal{E}\}|$. Accordingly, the degree matrix $\mathbf{D}$ of a graph $\mathcal{G}$ corresponds to the diagonal matrix with the vertex degrees $d_1, \ldots, d_n$ on the diagonal. The main tools exploited by the random graph model proposed here are the graph Laplacian matrices. Different graph Laplacians have been defined in the literature. In this work, we consistently use the *unnormalized graph Laplacian*, given by $\mathbf{L} = \mathbf{D} - \mathbf{A}$. Although alternative modeling options might be pursued based on normalized forms of the Laplacian (von Luxburg, 2007; Ng et al., 2001), the unnormalized form perfectly fits our goals, as we are going to see. A thorough analysis of alternative choices lies beyond the scope of the present work.

Some basic facts related to the unnormalized Laplacian matrix can be summarized as follows (Mohar, 1991):

Proposition 1 *The unnormalized graph Laplacian* $\mathbf{L}$ *of an undirected graph* $\mathcal{G}$ *has the following properties: (i)* $\mathbf{L}$ *is symmetric and positive semi-definite; (ii) the smallest eigenvalue of* $\mathbf{L}$ *is 0; (iii)* $\mathbf{L}$ *has* n *non-negative, real-valued eigenvalues* $0 = \lambda_1 \leq \ldots \leq \lambda_n$*; (iv) the multiplicity of the eigenvalue 0 of* $\mathbf{L}$ *equals the number of connected components in the graph, that is,* $\lambda_1 = 0$ *and* $\lambda_2 > 0$ *if and only if* $\mathcal{G}$ *is connected.*

In the following, the (algebraic) multiplicity of an eigenvalue λ_i will be denoted by $M(\lambda_i, \mathcal{G})$.

Based on property (*iv*) from Proposition 1, if we restrict our attention to connected graphs only, then the smallest non-zero eigenvalue is always given by $\lambda_2(\mathcal{G})$. This value is traditionally referred to as the *Fiedler eigenvalue*. The Fiedler eigenvalue provides insight into several graph properties. When there is a nontrivial spectral gap, i.e. $\lambda_2(\mathcal{G})$ is clearly separated from 0, the graph has good expansion properties, stronger connectivity, and rapid convergence of estimates based on random walks in the graph. Also, it is known that $\lambda_2(\mathcal{G}) \leq \mu(\mathcal{G})$, where $\mu(\mathcal{G})$ is the size of the smallest edge cut whose removal makes the graph disconnected (Mohar, 1991). Clearly, whenever the graph has more than one

connected component, then $\lambda_2(\mathcal{G})$ will be also equal to zero. However, in this work we abuse the term 'Fiedler eigenvalue' to denote the smallest eigenvalue different from zero, regardless of the number of connected components. That is, by Fiedler value we precisely mean the eigenvalue $\lambda_{k+1}(\mathcal{G})$, where $k = M(0, \mathcal{G})$.

For any pair of nodes u and v in a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, we define two corresponding graphs $\mathcal{G}^{uv^+}$ and $\mathcal{G}^{uv^-}$ in the following way: $\mathcal{G}^{uv^+} = (\mathcal{V}, \mathcal{E} \cup \{\{u, v\}\})$, and $\mathcal{G}^{uv^-} = (\mathcal{V}, \mathcal{E} \setminus \{\{u, v\}\})$. Clearly, we have that either $\mathcal{G}^{uv^+} = \mathcal{G}$ or $\mathcal{G}^{uv^-} = \mathcal{G}$. A basic property concerning the Laplacian eigenvalues of $\mathcal{G}^{uv^+}$ and $\mathcal{G}^{uv^-}$ is the following (Mohar, 1991; Anderson and Morley, 1985; Cvetković et al., 1979):

Lemma 1 *For any graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $|\mathcal{V}| = n$, we have that: (i) $\sum_{i=1}^{n} \lambda_i(\mathcal{G}^{uv^+}) - \lambda_i(\mathcal{G}^{uv^-}) = 2$; (ii) for any i such that $1 \leq i \leq n$, $\lambda_i(\mathcal{G}^{uv^-}) \leq \lambda_i(\mathcal{G}^{uv^+})$.*

4. Fiedler Random Fields

Fiedler random fields are introduced in Section 4.1, while in Section 4.2 we discuss their dependence structure.

4.1 Probability Distribution

Using the notions reviewed above, we define the Fiedler delta function $\Delta\lambda_2$ in the following way:

Definition 1 *Given graph $\mathcal{G}$, let $k = M(0, \mathcal{G}^{uv^+})$. Then,*

$$\Delta\lambda_2(u, v, \mathcal{G}) = \lambda_{k+1}(\mathcal{G}^{uv^+}) - \lambda_{k+1}(\mathcal{G}^{uv^-}) \tag{1}$$

In other words, for any pair of nodes u and v in graph $\mathcal{G}$, the Fiedler delta value of the pair $\{u, v\}$ in $\mathcal{G}$ is the (absolute) variation in the Fiedler eigenvalue of the graph Laplacian that would result from removing edge $\{u, v\}$ from $\mathcal{G}^{uv^+}$. To avoid possible confusion, it is useful to emphasize two points. First, although the Fiedler delta value of u and v in $\mathcal{G}$ is denoted by $\Delta\lambda_2(u, v, \mathcal{G})$, this does not entail that $\lambda_{k+1}(\mathcal{G}^{uv^+}) = \lambda_2(\mathcal{G}^{uv^+})$, i.e. that $M(0, \mathcal{G}^{uv^+}) = 1$. This is because we are interested in the smallest nonzero eigenvalue of the graph Laplacian, independent the multiplicity of the zero eigenvalue. In this respect, a less misleading notation for the Fiedler delta function would be $\Delta\lambda_{k+1}$, where k is defined as above, but to avoid clutter, we prefer the simpler notation $\Delta\lambda_2$. Second, it is not necessarily the case that $M(0, \mathcal{G}^{uv^+}) = M(0, \mathcal{G}^{uv^-})$, since removing edge $\{u, v\}$ from $\mathcal{G}^{uv^+}$ can increase the number of connected components in the graph. In such a case, we would have that $M(0, \mathcal{G}^{uv^-}) = M(0, \mathcal{G}^{uv^+}) + 1$ and $\lambda_{k+1}(\mathcal{G}^{uv^-}) = 0$, and (as a consequence) that $\Delta\lambda_2(u, v, \mathcal{G}) = \lambda_{k+1}(\mathcal{G}^{uv^+})$.

Concerning the range of the Fiedler delta function, we can easily prove the following proposition:

Proposition 2 *For any graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ and any pair of nodes $\{u, v\}$ such that $X_{uv} = 1$, we have that $0 \leq \Delta\lambda_2(u, v, \mathcal{G}) \leq 2$.*

Proof Let $k = M(0, \mathcal{G})$. The proposition follows straightforwardly from Lemma 1, given that $\Delta\lambda_2(u, v, \mathcal{G}) = \lambda_{k+1}(\mathcal{G}) - \lambda_{k+1}(\mathcal{G}^{uv^-})$. ■

We now proceed to define FRFs. Given a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, for each (unordered) pair of nodes $\{u, v\}$ such that $u \neq v$, we take X_{uv} to denote a binary random variable such that $X_{uv} = 1$ if $\{u, v\} \in \mathcal{E}$, and $X_{uv} = 0$ otherwise. Since the graph is undirected, $X_{uv} = X_{vu}$. We also say that a subgraph $\mathcal{G}_\mathcal{S}$ of $\mathcal{G}$ with edge set $\mathcal{E}_\mathcal{S}$ is incident to X_{uv} if $\{u, v\} \subseteq \bigcup_{e \in \mathcal{E}_\mathcal{S}} e$, i.e. if $\mathcal{G}_\mathcal{S}$ not only contains nodes u and v, but it also has at least one edge incident to u and one incident to v. To avoid confusion, notice that stating that $\{u, v\} \subseteq \bigcup_{e \in \mathcal{E}_\mathcal{S}} e$ is quite different from stating that $\{u, v\} \in \mathcal{E}_\mathcal{S}$, i.e. that the former statement is not a sufficient condition for u and v being linked by an edge. Then:

Definition 2 *Given a graph $\mathcal{G}$, let $\boldsymbol{X}_\mathcal{G}$ denote the set of random variables defined on $\mathcal{G}$, i.e. $\boldsymbol{X}_\mathcal{G} = \{X_{uv} \colon u \neq v \wedge \{u, v\} \subseteq \mathcal{V}\}$. For any $X_{uv} \in \boldsymbol{X}_\mathcal{G}$, let $\mathcal{G}_{uv}$ be a subgraph of $\mathcal{G}$ which is incident to X_{uv} and $\varphi_{uv} : \{0, 1\} \times [0, 2] \to \mathbb{R}$ be a function with parameter vector $\boldsymbol{\theta}$. We say that the probability distribution of $\boldsymbol{X}_\mathcal{G}$ is a* Fiedler random field *if it factorizes as*

$$P(\boldsymbol{X}_\mathcal{G} | \, \boldsymbol{\theta}) = \frac{1}{Z(\boldsymbol{\theta})} \exp \left(\sum_{X_{uv} \in \boldsymbol{X}_\mathcal{G}} \varphi_{uv}\big(X_{uv}, \Delta\lambda_2(u, v, \mathcal{G}_{uv}); \boldsymbol{\theta}\big) \right) \tag{2}$$

where $Z(\boldsymbol{\theta})$ is the partition function.

In other words, a FRF is a Gibbs-Boltzmann distribution over graphs, with potential functions defined for each node pair $\{u, v\}$ along with some neighboring subgraph $\mathcal{G}_{uv}$. In particular, in order to model the dependence of each variable X_{uv} on $\mathcal{G}_{uv}$, potentials take as argument both the value of X_{uv} and the Fiedler delta corresponding to $\{u, v\}$ in $\mathcal{G}_{uv}$. The idea is to treat the Fiedler delta statistic as a (real-valued) random variable defined over subgraph configurations, and to exploit this random variable as a compact representation of those configurations. This means that the dependence structure of a FRF is fixed by the particular choice of subgraphs $\mathcal{G}_{uv}$, so that the set $\boldsymbol{X}_{\mathcal{G}_{uv}} \setminus \{X_{uv}\}$ makes X_{uv} independent of $\boldsymbol{X}_\mathcal{G} \setminus \boldsymbol{X}_{\mathcal{G}_{uv}}$. Three fundamental questions are then the following. First, how do we fix the subgraph $\mathcal{G}_{uv}$ for each pair of nodes $\{u, v\}$? Second, how do we choose a shape for the potential functions, so as to fully exploit the information contained in the Fiedler delta, while avoiding unwarranted assumptions concerning their parametric form? Third, how does the Fiedler delta statistic behave with respect to the Markov dependence property for random graphs? One basic result related to the third question is presented in Section 4.2, while Section 5 will address the first two points.

4.2 Dependence Structure

We first recall the definition of Markov dependence for random graphs (Frank and Strauss, 1986). Let $\mathcal{N}(X_{uv})$ denote the set $\{X_{wz} \colon \{w, z\} \in \mathcal{E} \wedge |\{w, z\} \cap \{u, v\}| = 1\}$. We refer to $\mathcal{N}(X_{uv})$ as the *neighborhood* of X_{uv}. Then:

Definition 3 *A random graph $\mathcal{G}$ is said to be a* Markov graph *(or to have a* Markov dependence structure*) if, for any pair of variables X_{uv} and X_{wz} in $\mathcal{G}$ such that $\{u, v\} \cap \{w, z\} = \emptyset$, we have that $P(X_{uv} | \, X_{wz}, \mathcal{N}(X_{uv})) = P(X_{uv} | \, \mathcal{N}(X_{uv}))$.*

Based on Definition 3, we say that the dependence structure of a probability distribution over graphs is *non-Markovian* if, for disjoint pairs of nodes $\{u, v\}$ and $\{w, z\}$, it is consistent with the inequality $P(X_{uv}|\, X_{wz}, \mathcal{N}(X_{uv})) \neq P(X_{uv}|\mathcal{N}(X_{uv}))$. Informally, we have a non-Markovian dependence structure whenever considering the neighborhood of an edge variable X_{uv} is not enough to make X_{uv} independent of all the remaining variables in the graph. Concerning FRFs, we can prove the following proposition:

Proposition 3 *There exist Fiedler random fields with non-Markovian dependence structure.*

Proof Consider a FRF over a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ such that $\mathcal{V} = \{u, v, w, z\}$ and $\mathcal{E} = \{\{u, v\}, \{v, w\}, \{w, z\}, \{u, z\}\}$, and let $\mathcal{G}_{\mathcal{S}} = (\mathcal{V}_{\mathcal{S}}, \mathcal{E}_{\mathcal{S}})$ be the subgraph of $\mathcal{G}$ incident to $\mathcal{E}_{\mathcal{S}} = \mathcal{E} \backslash \{\{w, z\}\}$. Based on Def. 3, in order to show that $\mathcal{G}$ is non-Markovian it is sufficient to verify that $\mathcal{G}$ is consistent with the inequality $P(X_{uv}|\, X_{wz}, X_{vw}, X_{uz}) \neq P(X_{uv}|\, X_{vw}, X_{uz})$. By the Hammersley-Clifford theorem (Besag, 1974), this reduces to showing that $\mathcal{G}$ is consistent with $\varphi_{uv}(X_{uv}, \Delta\lambda_2(u, v, \mathcal{G}); \boldsymbol{\theta}) \neq \varphi_{uv}(X_{uv}, \Delta\lambda_2(u, v, \mathcal{G}_{\mathcal{S}}); \boldsymbol{\theta})$. We use the following result (Fiedler, 1973): if graphs $\mathcal{G}_1$ and $\mathcal{G}_2$ are, respectively, a path and a circuit of size n, then $\lambda_2(\mathcal{G}_1) = 2\,(1 - \cos(\pi/n))$ and $\lambda_2(\mathcal{G}_2) = 2\,(1 - \cos(2\pi/n))$. Since the configuration of $\mathcal{G}$ and $\mathcal{G}_{\mathcal{S}}$ is given by a circuit and a path respectively, where both have size 4, it follows that $\lambda_2(\mathcal{G}) = 2\,(1 - \cos(\pi/2))$ and $\lambda_2(\mathcal{G}_{\mathcal{S}}) = 2\,(1 - \cos(\pi/4))$. Also, we have that $\lambda_2(\mathcal{G}^{uv^-}) = \lambda_2(\mathcal{G}_{\mathcal{S}})$, since $\mathcal{G}^{uv^-}$ is also a path of size 4, and that $M(0, \mathcal{G}_{\mathcal{S}}^{uv^-}) = M(0, \mathcal{G}_{\mathcal{S}}) + 1$, since $\mathcal{G}_{\mathcal{S}}^{uv^-}$ has one more connected component than $\mathcal{G}_{\mathcal{S}}$. Therefore, $\Delta\lambda_2(u, v, \mathcal{G}) = 2\,\cos(\pi/4)$ and $\Delta\lambda_2(u, v, \mathcal{G}_{\mathcal{S}}) = 2\,(1 - \cos(\pi/4))$, i.e. $\Delta\lambda_2(u, v, \mathcal{G}) \neq \Delta\lambda_2(u, v, \mathcal{G}_{\mathcal{S}})$. Because of this inequality, there will exist parameterizations of φ_{uv} such that $\varphi_{uv}(X_{uv}, \Delta\lambda_2(u, v, \mathcal{G}); \boldsymbol{\theta}) \neq \varphi_{uv}(X_{uv}, \Delta\lambda_2(u, v, \mathcal{G}_{\mathcal{S}}); \boldsymbol{\theta})$, which means that the dependence structure of $\mathcal{G}$ is non-Markovian. ■

Proposition 3 is agnostic with respect to the relevance of non-Markovian dependence structures in real-world network modeling. We believe that the knowledge we have today concerning large-scale networks is not deep enough to precisely assess the importance of capturing such dependences when bulding statistical network models. On the other hand, the Markov independence property is an extremely popular tool for analyzing the independence structure of probabilistic models. For this reason, understanding how FRFs behave with respect to that property allows to easily relate this model to alternative statistical models (such as ERGs), where the analysis of Markov properties plays a central role in their mathematical formalization.

We stress the fact that Proposition 3 generally holds for the dependence between two variables X_{uv} and X_{wz} in circuits/paths of arbitrary size n, since the expression used for the Fiedler eigenvalues of such graphs holds for any n. In order to generalize the argument to the dependence between variables X_{uv} and X_{wz} in circuits/paths of arbitrary size, suppose that the 4-nodes circuit $\mathcal{G}$ used in the proof is replaced by a circuit $\mathcal{G}^* = (\mathcal{V}^*, \mathcal{E}^*)$ of size n, where $\mathcal{V}^* = \mathcal{V} \cup \{s_1, \ldots, s_1, \ldots, s_m, t_1, \ldots, t_m\}$ and $\mathcal{E}^*$ is obtained from $\mathcal{E}$ by replacing $\{u, z\}$ and $\{v, w\}$, respectively, with a path from u to z going through $s_1, \ldots, s_m$ and a path from v to w going through $t_1, \ldots, t_m$, so that $n = 2m+4$. In this case, if $\mathcal{G}^*_{\mathcal{S}}$ is the subgraph of $\mathcal{G}^*$ incident to $\mathcal{E}^*_{\mathcal{S}} = \mathcal{E}^* \setminus \{\{w, z\}\}$, we have again that $\Delta\lambda_2(u, v, \mathcal{G}^*) \neq \Delta\lambda_2(u, v, \mathcal{G}^*_{\mathcal{S}})$, which means that there exist FRFs such that $\varphi_{uv}(X_{uv}, \Delta\lambda_2(u, v, \mathcal{G}^*); \boldsymbol{\theta}) \neq \varphi_{uv}(X_{uv}, \Delta\lambda_2(u, v, \mathcal{G}^*_{\mathcal{S}}); \boldsymbol{\theta})$. This

fact suggests that FRFs allow to model edge correlations at virtually any distance within $\mathcal{G}$, provided that each subgraph $\mathcal{G}_{uv}$ is chosen in such a way as to encompass the relevant correlation.

5. Model Estimation

The problem of learning a FRF from an observed network can be split into the task of factorizing the overall joint distribution into a suitable set of subgraphs and the task of estimating the potential functions once a particular factorization has been fixed. The former task corresponds to estimating the dependence structure of the model, while we refer to the latter as a parameter estimation task. Here we develop one complete solution to the problem of learning the FRF potentials (Section 5.1), whereas we describe some heuristic ways to fix the dependence structure of the model (Section 5.2).

5.1 Parameter Learning

In order to estimate the FRF potentials, we need to specify on the one hand a suitable parametric form for such functions, and on the other hand the objective function that we want to optimize. The first task consists in fixing the range of possible values to be assigned to the parameter vector $\boldsymbol{\theta}$, as well as the particular form of the potential functions φ_{uv}. In order to make a good choice for the potentials architecture, we first need some insight into the behavior of the Fiedler delta statistic. Formally, the function $\Delta\lambda_2(u, v, \mathcal{G}_{uv})$ can be treated as a continuous random variable, with values in the closed interval $[0, 2]$. Such a variable will be distributed according to a specific density function $p\big(\Delta\lambda_2(u, v, \mathcal{G}_{uv})\big)$. Distinguishing linked and unlinked node pairs, we get the conditional densities $p\big(\Delta\lambda_2(u, v, \mathcal{G}_{uv})|\, X_{uv} = 1\big)$ and $p\big(\Delta\lambda_2(u, v, \mathcal{G}_{uv})|\, X_{uv} = 0\big)$, which we can denote by p_1 and p_0 respectively. Clearly, simple architectures for the potentials (such as linear functions) will be a realistic model only when the relative behavior of p_1 and p_0 takes a particularly simple form, e.g. when the respective data points are linearly separable. Otherwise, linear potentials will be a poor fit to the observed data, and a more complex architecture will be needed. As a matter of fact, the two densities can assume a variety of shapes in different settings, hence violating any simplistic expectation we may have a priori. The phenomenon is illustrated in Figure 1 over four different networks. For example, notice how the density p_1 can be (roughly) considered as unimodal for the small-world network, whereas it seems to have at least two modes for preferential attachment and three or more in the scientific collaboration and the protein-protein interaction networks. Such observations lead us to think that any simplifying assumption concerning the shape of such densities could severely limit our capability of estimating the potential functions with reasonable accuracy.

Therefore, we choose to model potential functions by a feed-forward multilayer perceptron (MLP), due to its well-known capabilities of approximating functions of arbitrary shape (Cybenko, 1989; Hornik, 1991). Throughout the applications described in this paper we use a standard MLP architecture with one hidden layer and hyperbolic tangent activation functions. Therefore, our vector $\boldsymbol{\theta}$ simply consists of the weights specified for our MLP estimator, where the resulting function is denoted by φ. As it happens, preliminary investigation revealed linear potential functions to be an extremely poor model for anything but the simplest sorts of networks. Notice that, as far as the estimation of potentials is

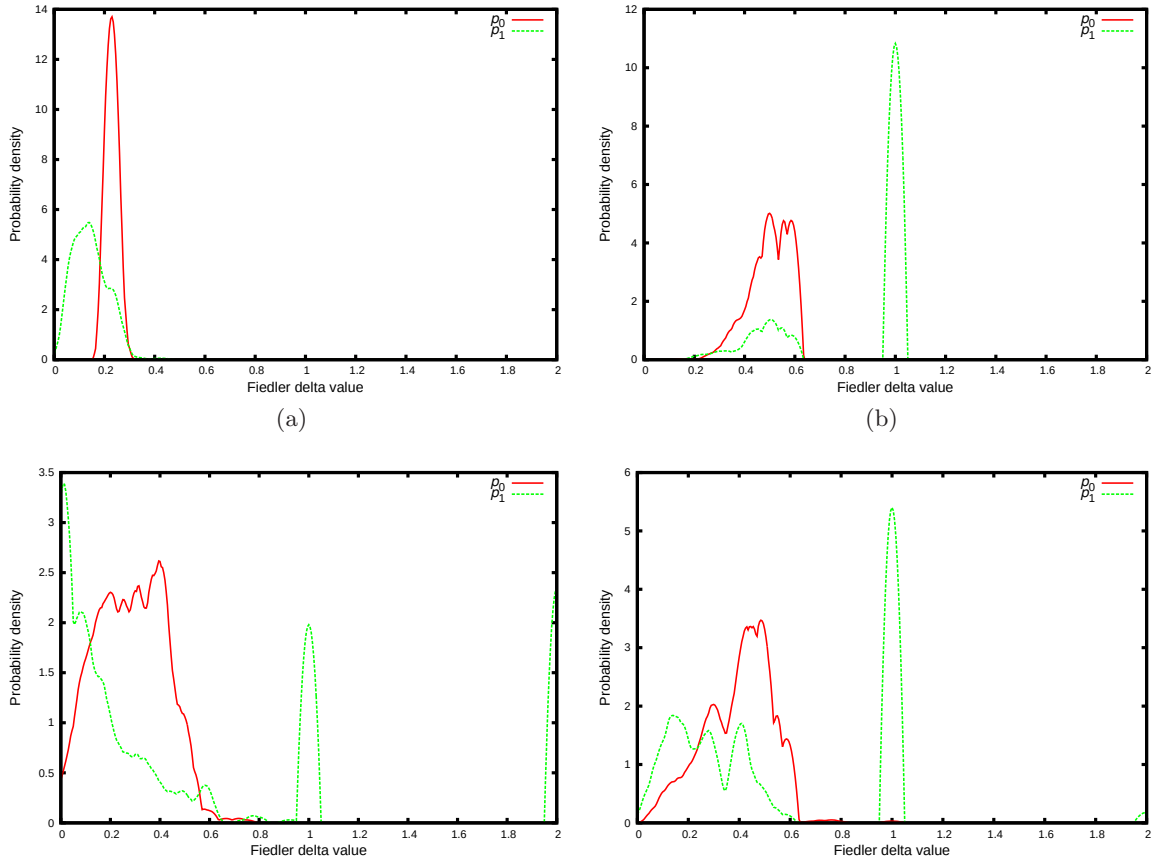


Figure 1: Estimates of the density functions $p_0(\Delta\lambda_2(u, v, \mathcal{G}_{uv}))$ and $p_1(\Delta\lambda_2(u, v, \mathcal{G}_{uv}))$ over four different networks: small-world (a), preferential attachment (b), scientific collaboration (c), and protein-protein interaction (d). Subgraphs are drawn using one-wave snowball sampling. Plots (a)–(d) correspond to the networks described in Section 6 as Synth-WS, Synth-BA, CA-HepPh, and PPI-Drosoph respectively.

concerned, any regression model offering approximation capabilities analogous to the MLP family could be used as well. Here, the only requirement is to avoid unwarranted prior assumptions with respect to the shape of the potential functions. In this respect, we take our approach to be genuinely *nonparametric*, since it does not require the parametric form of the target functions to be known a priori in order to estimate them accurately.

Concerning instead the learning objective, the main difficulty we want to avoid is the complexity of computing the partition function involved in the Gibbs-Boltzmann distribution. The approach we adopt to this aim is to minimize a *contrastive divergence* objective (Hinton, 2002). If $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ is the network that we want to fit our model to, and $\mathcal{G}_{uv} = (\mathcal{V}_{uv}, \mathcal{E}_{uv})$ is a subgraph of $\mathcal{G}$ such that $\{u, v\} \subseteq \mathcal{V}_{uv}$, let $\mathcal{G}^*_{uv}$ denote the graph that we obtain by resampling the value of X_{uv} in $\mathcal{G}_{uv}$ according to the conditional distribution

$\widehat{P}(X_{uv}|\, \boldsymbol{x}_{\mathcal{G}_{uv}} \setminus \{x_{uv}\}; \boldsymbol{\theta})$ predicted by our model. In other words, $\mathcal{G}^*_{uv}$ is the result of performing just one iteration of Gibbs sampling on X_{uv} using $\boldsymbol{\theta}$, where the configuration $\boldsymbol{x}_{\mathcal{G}_{uv}}$ of $\mathcal{G}_{uv}$ is used to initialize the (single-step) Markov chain. Then, our goal is to minimize the function $\ell_{CD}(\boldsymbol{\theta}; \mathcal{G})$, given by:

$$\begin{aligned} \ell_{CD}(\boldsymbol{\theta}; \mathcal{G}) &= \log \left\{ \frac{1}{Z(\boldsymbol{\theta})} \exp \left(\sum_{X_{uv} \in \boldsymbol{X}_{\mathcal{G}}} \varphi\big(x^*_{uv}, \Delta\lambda_2(u, v, \mathcal{G}^*_{uv}); \boldsymbol{\theta}\big) \right) \right\} - \log \widehat{P}(\boldsymbol{x}_{\mathcal{G}}|\, \boldsymbol{\theta}) \\ &= \sum_{X_{uv} \in \boldsymbol{X}_{\mathcal{G}}} \left\{ \varphi\big(x^*_{uv}, \Delta\lambda_2(u, v, \mathcal{G}^*_{uv}); \boldsymbol{\theta}\big) - \varphi\big(x_{uv}, \Delta\lambda_2(u, v, \mathcal{G}_{uv}); \boldsymbol{\theta}\big) \right\} \end{aligned} \tag{3}$$

where φ is the function computed by our MLP architecture. The appeal of contrastive divergence learning is that, while it does not require to compute the partition function, it is known to converge to points which are very close to maximum-likelihood solutions (Á. Carreira-Perpiñán and Hinton, 2005). In practice, if we want our learning objective to be usable in the large-scale setting, then it is not feasible to sum over all node pairs $\{u, v\}$ in the network, since the number of such pairs grows quadratically with $|\mathcal{V}|$. In this respect, a straightforward approach for scaling to very large networks consists in sampling n objects from the set of all possible pairs of nodes, taking care that the sample contains a good balance between linked and unlinked pairs. Once sampled our training set $\mathcal{D} = \big\{(x_{u_1v_1}, \mathcal{G}_{u_1v_1}), \ldots, (x_{u_nv_n}, \mathcal{G}_{u_nv_n})\big\}$, we learn the MLP weights by minimizing the objective $\ell_{CD}(\boldsymbol{\theta}; \mathcal{D})$, which we obtain from $\ell_{CD}(\boldsymbol{\theta}; \mathcal{G})$ by restricting the summation in Equation 3 to the elements of $\mathcal{D}$. In our applications, minimization is performed by iterative gradient descent, using standard backpropagation for updating the MLP weights.

5.2 Structure Learning

Another issue we need to address concerns the way we sample a suitable set of subgraphs $\mathcal{G}_{u_1v_1}, \ldots, \mathcal{G}_{u_nv_n}$ for the selected pairs of nodes. Although different sampling techniques could be used in principle (Leskovec and Faloutsos, 2006), our goal is to model correlations between each variable X_{uv} and some neighboring region $\mathcal{G}_{uv}$ in $\mathcal{G}$. Such a neighborhood should be large enough to make $\Delta\lambda_2(u, v, \mathcal{G}_{uv})$ sufficiently informative with respect to the overall network, but also small enough to keep the spectral decomposition of $\mathcal{G}_{uv}$ computationally tractable. Here, in order to sample $\mathcal{G}_{uv}$, we propose three different methods for drawing $\mathcal{V}_{uv}$, using u and v as seeds. Once the nodes have been sampled, we set $\mathcal{E}_{uv}$ to be the edge set *induced* by $\mathcal{V}_{uv}$ in $\mathcal{G}$, that is $\mathcal{E}_{uv} = \{e \in \mathcal{E} \colon e \subseteq \mathcal{V}_{uv}\}$.

The first option we suggest is a variant of random walk sampling (Leskovec and Faloutsos, 2006). Given nodes u and v, we perform two random walks in parallel, starting from u and v respectively. At every step of each random walk, we jump back to the starting node with probability P, otherwise we proceed by sampling (uniformly at random) one neighbor of the currently visited node. The two walks stop when the number of sampled nodes reaches some specified value S, or when the connected components containing u and v have been explored exhaustively without reaching the desired sample size, whatever condition is met first (see Algorithm 1).

An alternative approach which is well suited to our setting is snowball sampling (Kolaczyk, 2009). Here, we perform k sampling 'waves' on $\mathcal{G}$, starting from u and v as seeds.

Algorithm 1 `RandomWalkSample`: Sampling a random neighboring subgraph for a given pair of nodes

Input: Undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$; node pair $\{u, v\}$; size S of the sample to draw; probability P of jumping back to u or v.
Output: Undirected graph $\mathcal{G}_{uv} = (\mathcal{V}_{uv}, \mathcal{E}_{uv})$.

`RandomWalkSample`$(\mathcal{G}, \{u, v\}, S, P)$:
1. $\mathcal{V}_u$ = $\{w \in \mathcal{V}$: w is in the same connected component as $u\}$
2. $\mathcal{V}_v$ = $\{w \in \mathcal{V}$: w is in the same connected component as $v\}$
3. `if`$(|\mathcal{V}_u \cup \mathcal{V}_v| \leq S)$
4. $\quad \mathcal{V}_{uv}$ = $\mathcal{V}_u \cup \mathcal{V}_v$
5. `else`
6. $\quad \mathcal{V}_{uv}$ = $\{u, v\}$
7. $\quad w_u$ = u
8. $\quad w_v$ = v
9. $\quad$ `while`$(|\mathcal{V}_{uv}| < S)$
10. $\quad\quad$ `if`$(d_{w_u} > 0)$
11. $\quad\quad\quad p$ = random real in $[0, 1)$
12. $\quad\quad\quad$ `if`$(p < P)$
13. $\quad\quad\quad\quad w_u$ = u
14. $\quad\quad\quad$ `else`
15. $\quad\quad\quad\quad w_u$ = random node in $\mathcal{N}_{w_u}$
16. $\quad\quad$ `if`$(d_{w_v} > 0)$
17. $\quad\quad\quad p$ = random real in $[0, 1)$
18. $\quad\quad\quad$ `if`$(p < P)$
19. $\quad\quad\quad\quad w_v$ = v
20. $\quad\quad\quad$ `else`
21. $\quad\quad\quad\quad w_v$ = random node in $\mathcal{N}_{w_v}$
22. $\quad\quad \mathcal{V}_{uv}$ = $\mathcal{V}_{uv} \cup \{w_u, w_v\}$
23. $\mathcal{E}_{uv}$ = $\{\{w, z\} \in \mathcal{E}$: $\{w, z\} \subseteq \mathcal{V}_{uv}\}$
24. `return` $(\mathcal{V}_{uv}, \mathcal{E}_{uv})$

Each wave consists in expanding the set of currently sampled nodes by adding to it all nodes that are adjacent to at least one element of the set (see Algorithm 2).

As a third sampling option, we propose to prune snowball subgraphs according to the following criterion. Given the subgraph sample returned by k snowball waves for the seeds u and v, any other node w from that sample is deleted if it does not lie on at least one path connecting u and v within the subgraph. Intuitively, since the Fiedler delta value for nodes u and v measures the impact of having u and v linked in a given subgraph, the proposed criterion has the effect of restricting the computation of $\Delta\lambda_2(u, v, \mathcal{G}_{uv})$ to only those nodes in $\mathcal{G}_{uv}$ that actually affect the connectivity structure linking u and v, i.e. that allow information to flow between u and v. We refer to such subgraph samples as *snowball frames* (see Algorithm 3).

Algorithm 2 `SnowballSample`: Sampling a snowball subgraph for a given pair of seeds

Input: Undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$; node pair $\{u, v\}$; number k of snowball waves.
Output: Undirected graph $\mathcal{G}_{uv} = (\mathcal{V}_{uv}, \mathcal{E}_{uv})$.

`SnowballSample`$(\mathcal{G}, \{u, v\}, k)$:
1. $\mathcal{V}_{uv}$ = $\{u, v\}$
2. for(i = 1 to k)
3. $\quad \mathcal{V}_{uv} = \mathcal{V}_{uv} \cup \bigcup_{w \in \mathcal{V}_{uv}} \{z \in \mathcal{V} : \{w, z\} \in \mathcal{E}\}$
4. $\mathcal{E}_{uv} = \{\{w, z\} \in \mathcal{E} : \{w, z\} \subseteq \mathcal{V}_{uv}\}$
5. return $(\mathcal{V}_{uv}, \mathcal{E}_{uv})$

Algorithm 3 `SnowballFrameSample`: Sampling a snowball frame for a given pair of seeds

Input: Undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$; node pair $\{u, v\}$; number k of snowball waves.
Output: Undirected graph $\mathcal{G}_{uv} = (\mathcal{V}_{uv}, \mathcal{E}_{uv})$.

`SnowballFrameSample`$(\mathcal{G}, \{u, v\}, k)$:
1. $(\mathcal{V}_{uv}, \mathcal{E}_{uv})$ = `SnowballSample`$(\mathcal{G}, \{u, v\}, k)$
2. for($w \in \mathcal{V}_{uv}$)
3. $\quad$ if(there exists no path from u to v in $(\mathcal{V}_{uv}, \mathcal{E}_{uv})$ going through w)
4. $\quad\quad \mathcal{V}_{uv} = \mathcal{V}_{uv} \setminus \{w\}$
5. $\mathcal{E}_{uv} = \{\{w, z\} \in \mathcal{E} : \{w, z\} \subseteq \mathcal{V}_{uv}\}$
6. return $(\mathcal{V}_{uv}, \mathcal{E}_{uv})$

Different sampling approaches may prove to be suitable with respect to different goals. In fact, the issue of evaluating the strengths and weaknesses of subgraph sampling algorithms is a relatively recent research topic, quite open both to theoretical and experimental contributions (Leskovec and Faloutsos, 2006; Hübler et al., 2008; Kolaczyk, 2009). While such an evaluation lies beyond the scope of the present work, our main interest consists in answering two questions. First, how large an impact has the chosen sampling algorithm on the overall accuracy of FRF estimation? Second, is it possible to assess the modeling capabilities and predictive power of FRFs while abstracting from the chosen sampling algorithm? With respect to these questions, our goal is to characterize at least a basic range of modeling options such that, even after switching from a given sampling method to a different one, FRF-based estimates will ensure a relatively stable prediction accuracy as compared to alternative statistical models. As already mentioned before, our approach to the structure learning problem mainly consists in developing a basic range of heuristic solutions and analyzing their strengths and weaknesses. A complete treatment of this problem goes clearly beyond the scope of the present contribution.

The performance of Algorithms 1–3 in FRF learning will be analyzed experimentally in Section 6. One clear difference between random walk sampling and the snowball methods is that the former allows a much stricter control of subgraph size, which is warranted not to exceed the specified S value. With snowball-based sampling it is instead more difficult

to control sample size, since its expected value will depend not simply on the specified number of waves, but mainly on the degree distribution of the given network. Yet, snowball frames will typically have a smaller size than their snowball counterparts, because of the involved pruning process. Other things being equal, a small subgraph size will be generally desirable because of reducing the complexity of eigendecomposition, which is cubic with respect to that size (Bai et al., 2000). On the other hand, Algorithms 2–3 are more suitable for controlling the maximum span of statistical dependence between non-incident edges, i.e. the maximum distance between edge pairs which we are willing to consider jointly in the computation of potentials. Such a constraint on the range of considered edge correlations cannot be imposed easily through random walk sampling, since that range is only indirectly (and uncertainly) affected by the specified sample size.

6. Experimental Analysis

In order to assess the performance of FRFs as models for large-scale networks, we design two different groups of experiments, in link prediction and graph generation, described in Sections 6.2 and 6.3 respectively. Before presenting such experiments, however, it will be useful to analyze the Fiedler delta statistic as such, in order to gain some hands-on understanding of its empirical behavior. This point (which is addressed in Section 6.1) will be especially helpful in guiding several key choices in the design and implementation of FRFs. All the experiments are performed both on synthetic and on real-world networks, whose main features are summarized in Table 1. The artificial data comprise an ER, a WS, and a BA network. In the ER network, the prior probability of observing an edge is 10^{-4}. The WS network has average node degree $2\delta = 6$ and rewiring probability $\beta = 0.1$ (Watts and Strogatz, 1998), whereas for the BA network we set $\alpha = 0.5$ (Albert and Barabási, 2002). Although alternative network models might be considered in order to generate artificial data, we choose these three models because their extremely wide diffusion in the research community has made them a classic reference in network science. The point of our choice for the synthetic benchmarks is to focus not on the most realistic models (as on top of them we consider several real-world networks as well), but rather on models whose properties have been extensively investigated and universally understood. The real-world datasets include five social, one technological, and two biological networks. Social data (Leskovec et al., 2007) are collaboration networks drawn from the arXiv e-print repository (`http://snap.stanford.edu/data/`), where nodes represent scientists and edges represent paper coauthorships. The Skitter network (Leskovec et al., 2005) is an autonomous system, representing the internet topology extracted from a set of traceroutes in 2005 (`http://www.caida.org/tools/measurement/skitter/`). Finally, the two biological datasets consist in protein-protein interaction networks observed in fruit flies and yeast respectively (`http://www.comp.nus.edu.sg/~whsu/IRAP/datasets.html`). All of the models and techniques considered in the experimental evaluation are implemented in the JProGraM open-source library (`http://jprogram.sourceforge.net/`), except for stochastic Kronecker graphs, which we use through the official SNAP implementation (`http://snap.stanford.edu/snap/`). In order to ease replication of the experiments, we also make available all the training/test datasets mentioned below at `http://researchers.lille.inria.fr/~freno/datasets.html`.

Network	$\lvert\mathcal{V}\rvert$	$\lvert\mathcal{E}\rvert$	$CC_{\mathcal{G}}$	$D_{\mathcal{G}}$	$P(X_{uv}=1)$
Synth-ER	50,000	124,883	0.00	14	$9 \cdot 10^{-5}$
Synth-WS	50,000	150,000	0.44	17	10^{-4}
Synth-BA	50,000	49,999	0.00	36	$4 \cdot 10^{-5}$
CA-AstroPh	18,772	396,160	0.63	14	$2 \cdot 10^{-3}$
CA-CondMat	23,133	186,936	0.63	15	$6 \cdot 10^{-4}$
CA-GrQc	5,242	28,980	0.52	17	$2 \cdot 10^{-3}$
CA-HepPh	12,008	237,010	0.61	13	$3 \cdot 10^{-3}$
CA-HepTh	9,877	51,971	0.47	17	10^{-3}
AS-Skitter	1,696,415	11,095,298	0.29	25	$7 \cdot 10^{-6}$
PPI-Drosoph	7,679	22,563	0.04	11	$7 \cdot 10^{-4}$
PPI-Sacchar	4,136	7,098	0.06	14	$8 \cdot 10^{-4}$

Table 1: General statistics for the networks used in the experiments. $CC_{\mathcal{G}}$, $D_{\mathcal{G}}$, and $P(X_{uv}=1)$ denote the average clustering coefficient, the network diameter, and the marginal edge probability respectively.

6.1 Density of the Fiedler Delta Value

In this block of experiments, our goal is to gain some insight into the empirical behavior of the Fiedler delta statistic. Since the Fiedler delta is the basic feature on top of which our random field model is built, exploring the behavior of its density provides a transparent way of assessing its representational capabilities with respect to different kinds of networks and different subgraph sampling approaches. Virtually, the more information is enclosed in the Fiedler delta function with respect to the presence/absence of edges *given the neighboring subgraphs*, the farther we should be able to go with FRFs in modeling probability distributions over networks.

To explore the behavior of the density $p\big(\Delta\lambda_2(u,v,\mathcal{G}_{uv})\big)$, we perform two kinds of measurements. On the one hand, let us divide all node pairs from $\mathcal{G}=(\mathcal{V},\mathcal{E})$ in two different classes, containing linked pairs (i.e. elements of $\mathcal{E}$) and unlinked pairs respectively. Then, two different versions can be defined for the Fiedler delta density, conditioning on linked and unlinked node pairs, that is $p\big(\Delta\lambda_2(u,v,\mathcal{G}_{uv})|\, X_{uv}=1\big)$ and $p\big(\Delta\lambda_2(u,v,\mathcal{G}_{uv})|\, X_{uv}=0\big)$. Let us denote the two density functions by p_1 and p_0 respectively. For these two densities we measure the Kolmogorov-Smirnov D-statistic, given by:

$$D(p_1,p_0)=\max_{0\leq x\leq 2}\big|F_1(x)-F_0(x)\big| \tag{4}$$

where each F_i is the cumulative distribution function corresponding to p_i. Informally, $D(p_1,p_0)$ expresses the extent to which the two density functions differ from one another. The higher the value of the D-statistic, the larger the distance between p_0 and p_1. Ideally, we want that value to be as high as possible, since intuitively it measures the discriminative capability of $\Delta\lambda_2$ with respect to linked and unlinked node pairs. In particular, when the D-statistic happens to fall below the relevant critical value, the Fiedler delta statistic will not be of any use at all in discriminating present edges from absent ones.

On the other hand, if we focus on the random variable $\Delta\lambda_2(u, v, \mathcal{G}_{uv})$ as such, we can use the mutual information criterion to measure the strength of the statistical correlation between the Fiedler delta and the edge variable X_{uv}. The mutual information of $\Delta\lambda_2(u, v, \mathcal{G}_{uv})$ and X_{uv} is given by the following quantity:

$$I\big(\Delta\lambda_2(u, v, \mathcal{G}_{uv}), X_{uv}\big) = \sum_{x_{uv}} \int_0^2 p(t|\, x_{uv}) P(x_{uv}) \log_2 \left(\frac{p(t|\, x_{uv})}{\sum_{x^*_{uv}} p(t|\, x^*_{uv}) P(x^*_{uv})} \right) \mathrm{d}t \qquad (5)$$

where $p(t|\, X_{uv})$ is the conditional density of $\Delta\lambda_2$ given the edge variable X_{uv}, and $P(X_{uv})$ is the marginal distribution of the latter. The higher the value of $I\big(\Delta\lambda_2(u, v, \mathcal{G}_{uv}), X_{uv}\big)$, the more useful will be the information conveyed by the Fiedler delta function for predicting the presence or absence of edges. As compared to the Kolmogorov-Smirnov D-statistic, mutual information can be interpreted as a more direct measure of the predictive power of $\Delta\lambda_2(u, v, \mathcal{G}_{uv})$ with respect to X_{uv}. On the other hand, the latter test does not provide any reference thresholds such as the critical values that come with the former. Therefore, the D-statistic and the mutual information criterion can be viewed as complementary sources of evidence concerning the discriminative power of the Fiedler delta statistic with respect to the presence/absence of links.

Both for the Kolmogorov-Smirnov and the mutual information test, we estimate density functions using Parzen windows (Rosenblatt, 1956; Parzen, 1962). Since the Fiedler delta density has a finite support, the Epanechnikov kernel is an appropriate choice in this case (Epanechnikov, 1969). In our implementation, the integrals from Equations 4–5 are approximated using the composite Simpson rule (Atkinson, 1989).

Results are plotted in Figures 2–5. For each dataset, we compare the three subgraph sampling algorithms described in Section 5.2, namely random walk (RW), snowball, and snowball frame (SB-k and SBF-k respectively, where k is the number of waves). D-statistic and mutual information values are plotted against a growing number of nodes in the sampled subgraphs. The different behavior of random walk and snowball-based algorithms with respect to the subgraph size makes it necessary to adopt different solutions in order to plot their results within the same reference frame. For the RW algorithm, D-statistic and mutual information values can be plotted straightforwardly as a function of the chosen subgraph size S. For the chosen S, we simply draw a suitable number n of subgraphs from the given network (where each subgraph has exactly S nodes), we estimate the relevant densities, and then we report the resulting value for the D-statistic or the mutual information test. On the other hand, in SB-k and SBF-k sampling, subgraph size depends only indirectly on the chosen number of waves. Therefore, after choosing a value for k, we sample n subgraphs from the network, where the number of nodes in each subgraph may vary arbitrarily. Once we estimate the Fiedler delta densities for the sampled datasets, we use the measured D-statistic/mutual information value as the y-coordinate of each data point (i.e. each subgraph sample), where the x-coordinate is given instead by its size. Consequently, for the snowball-based algorithms, the plot also provides information concerning the distribution of subgraph size in the sampled dataset.

The plotted results lend themselves to a number of considerations:

1. For all networks except the Synth-ER one, the D-statistic stays consistently above its critical value over a significant part of the considered range. This means that genuine

statistical dependence between edges and their neighboring subgraphs is present in all of the considered networks (except the one generated from a distribution which is explicitly assuming independence), which justifies our interest in modeling the joint distribution of network edges while limiting the involved independence assumptions;

2. The highest predictive values of the Fiedler delta statistic are achieved for subgraph sizes which do not seem to depend at all on the overall network size. This suggests that the relevant statistical correlations over collections of edges are inherently local, i.e. they tend to emerge at a very small scale within the network. Such an observation is equally supported by RW and SB/SBF plots;

3. For most real networks, both statistical tests exhibit decaying values for a growing subgraph size, typically falling below the critical value in the right-hand side of the considered interval. If we look at the synthetic data, this trend is clearly observed in the small-world network (Synth-WS), but not in the scale-free one (Synth-BA). In the latter case, the two tests stabilize instead once the subgraph size exceeds 100 nodes approximately, with the D-statistic remaining well above the critical value. This sort of stabilization described for the Synth-BA network is remarkably displayed by the protein-protein interaction data, which virtually sets them apart from the coauthorship and the autonomous system networks. While a deep understanding of this result goes beyond the present contribution, we stress how the common assumption that all these network families are analogous, in that they all obey power-law distributions, appears as a quite limiting approach to data analysis once we see how a different sort of statistic can bring to light properties which are not captured by the scale-free model. Significantly, more and more evidence is being put forward in the literature concerning the limitations of power-law models (Clauset et al., 2009);

4. On average, at least one of the snowball-based approaches is delivering better subgraph samples than the random walk approach. Moreover, one-wave snowball sampling is consistently better than two-waves sampling (for both the SB and the SBF variant), which is quite coherent with what happens to RW sampling as the subgraph size gets larger. On the other hand, neither SB nor SBF can be considered individually better than RW, since the relative performance of the three algorithms seems to display different patterns in different datasets.

6.2 Link Prediction

In this group of experiments, given a random network $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, our goal is to measure the accuracy of FRFs at estimating the conditional distribution of variables X_{uv} given the configuration of neighboring subgraphs $\mathcal{G}_{uv}$ of $\mathcal{G}$. This can be seen as a link prediction problem where only local information (given by $\mathcal{G}_{uv}$) can be used for predicting the presence of a link $\{u, v\}$. Recall that FRFs are trained on a data sample $\mathcal{D} = \left\{(x_{u_1v_1}, \mathcal{G}_{u_1v_1}), \ldots, (x_{u_nv_n}, \mathcal{G}_{u_nv_n})\right\}$. Let us sample our training set $\mathcal{D}$ by first drawing n node pairs from $\mathcal{V}$ in such a way that linked and unlinked pairs from $\mathcal{G}$ are equally represented in $\mathcal{D}$, and then extracting the corresponding subgraphs $\mathcal{G}_{u_i,v_i}$ by one of the sampling algorithms described in Section 5.2. We then learn the model parameters as described in Section 5.1. A test set $\mathcal{T}$ containing m objects $(x_{u_1v_1}, \mathcal{G}_{u_1v_1}), \ldots, (x_{u_mv_m}, \mathcal{G}_{u_mv_m})$ is also

sampled from $\mathcal{G}$ so that $\mathcal{T} \cap \mathcal{D} = \emptyset$, where pairs $\{u_i, v_i\}$ in $\mathcal{T}$ are again uniformly divided into linked and unlinked pairs. In all the experiments reported in this work, the number of hidden units in the MLP architecture is set to 5, where this value was tuned by preliminary exploration of the model behavior on validation data. Predictions are derived from the learned model by first computing the conditional probability of observing a link for each pair of nodes $\{u_j, v_j\}$ in $\mathcal{T}$, and then making a decision on the presence/absence of links by thresholding the predicted probability (where the threshold is tuned by cross-validation). Prediction accuracy over $\mathcal{T}$ is measured by averaging the accuracy values for linked and unlinked pairs, and repeating the measurement according to a 5-fold cross-validation scheme, where $|\mathcal{D}| = 800$ and $|\mathcal{T}| = 200$.

The key question to be answered concerns the usefulness of the dependence structure modeled by FRFs in predicting the conditional distributions of edges given their neighboring subgraphs. That is, we want to ascertain whether the effort of modeling the conditional independence structure of the overall network is justified by a suitable gain in prediction accuracy with respect to statistical models that do not focus explicitly on such dependence structure. Depending on which one of the different subgraph sampling algorithms we choose, we consequently obtain a different sort of FRF model, which can be more or less appropriate to each considered network depending on the statistical properties of the data at hand. To gain some insight into the variability of the behavior of FRFs as a function of the chosen sampling method, we show results for all three versions of the model, which we refer to as FRF-RW, FRF-SB, and FRF-SBF depending on whether we use Algorithm 1, 2, or 3 respectively. Both the number of nodes in RW sampling and the number of waves in SB and SBF sampling are tuned by preliminary experiments on validation data, leading to a general choice of 10 nodes and 1 wave respectively. Interestingly, such hyperparameter settings are quite coherent with the more general results of the experiments reported in Section 6.1. As considered already in Section 5.2, our goal is not to establish the superiority of anyone of the sampling algorithms, but to ascertain instead the robustness of FRF modeling with respect to the sampling approach. Ideally, we wish the performance of FRFs to maintain a relatively good level across those algorithms if compared to statistical models from different families. To this aim, we compare FRFs to several statistical models for large-scale networks, namely the WS and BA models, as well as MRGs and HRGs from the exponential family (see Section 2). We estimate the parameters of these models by (batch) gradient descent, conforming to previous usage (Freno et al., 2012). The ER model is not considered in this group of experiments, since the involved independence assumption makes it unusable (i.e. equivalent to random guessing) for the purposes of conditional estimation tasks. On the other hand, SKGs cannot be included either in this comparison (whereas they will be taken into account in the experiments on graph generation), since no expression is known in their case for conditional edge distributions.

Accuracy values for the different models are reported in Figure 6. Overall, FRFs regularly outperform all other models in each one of the considered networks, where a sanity check on the Synth-ER network shows instead a collapse of all approaches onto the baseline of mere random guessing. In particular, as the BA and WS model are almost always worse than the ERG- and FRF-based approaches, the results suggest that an explicit focus on dependence modeling offers higher predictive power. In other words, all networks (except the ER one) exhibit genuine dependences among edges, which is crucial in estimating their

distributions. At the same time, the accuracy gain of FRFs over ERGs seems to show that the sort of potential functions employed in the latter class of models do not offer a predictive power equivalent to what we observe for the spectral approach pursued in FRFs. This indicates that, on top of the focus on conditional independence estimation, the Laplacian spectrum is a promising place to consider when searching for the relevant edge correlations.

An additional point that we want to address with respect to FRF estimation concerns whether the overall network size (in terms of the number of nodes) has an impact on the number of training examples that will be necessary for FRFs to converge to stable prediction accuracy. Since FRFs are trained on a data sample $\mathcal{D} = \{(x_{u_1v_1}, \mathcal{G}_{u_1v_1}), \ldots, (x_{u_nv_n}, \mathcal{G}_{u_nv_n})\}$, where $n \ll \frac{|\mathcal{V}|(|\mathcal{V}|-1)}{2}$, converging to stable predictions for values of n which do not depend on $|\mathcal{V}|$ is a crucial requirement for achieving large-scale applicability. In Figure 7, the accuracy of FRFs on the test set $\mathcal{T}$ is plotted against a growing size of the training set $\mathcal{D}$, where $12 \leq |\mathcal{D}| \leq 48$ and $|\mathcal{T}| = 10,000$. To exclude any variability due to the network domain and the subgraph sampling approach, we restrict the focus to the arXiv coauthorship data, and we consistently sample the subgraphs using one snowball wave.

Interestingly, the number of training examples required for the accuracy curve to stabilize does not seem to depend at all on the overall network size. Indeed, fastest convergence is achieved for the average-sized and the second largest networks, i.e. CA-HepPh and CA-AstroPh respectively. Notice how a training sample containing an extremely small percentage of node pairs is sufficient for our learning approach to converge to stable prediction accuracy. This result encourages to think of FRFs as a very promising network model for the large-scale setting.

6.3 Graph Generation

For the purposes of data exploration and visualization, it is often desirable to generate artificial network samples which can serve as small-scale representations of a target, large-scale network. Therefore, what we want to assess now is whether the FRFs learned on different sorts of networks can be effectively used to generate small-scale artificial graphs mimicking some representative target properties from the real network, such as degree distribution (DD) and clustering coefficient distribution (CC). To this aim, we use Gibbs sampling to generate artificial graphs from the estimated FRF models, and then we compare the DD and CC observed in the artificial graphs with those estimated on the whole networks. We compare the graphs generated by FRFs to those generated by all of the models already considered in Section 6.2, as well as to SKGs. The distance in DD and CC distribution between the artificial graphs on the one hand and the corresponding real network on the other hand is measured using the Kolmogorov-Smirnov D-statistic, following a common use in network mining research (Leskovec and Faloutsos, 2006). In this case, lower values indicate more accurate results, because they suggest a stronger similarity between the generated graphs and the target network. The results are displayed in Figure 8. Values are averaged over 100 samples, where each sample contains 128 nodes.

The outcome motivates the following considerations. FRFs (in at least one of their variants) and SKGs are the best modeling options everywhere, both for DD and CC, with the BA model being typically very close to their performance. On the other hand, ERG models deliver fairly inaccurate results, on a par with WS samples. If we look at these re-

sults together with the link prediction experiments, we can also make the following remark. Excluding FRFs, we see that in link prediction, among the available models, the ones explicitly focusing on conditional independence estimation (namely ERGs) achieve better results than models which are not designed to capture dependence patterns over edges (i.e. BA and WS). On the contrary, the former models are much less accurate than BA and especially than SKGs when the goal is to mimick high-level network properties through small-scale representations. At first glance, this might suggest that the challenges of conditional independence modeling and large-scale graph summarization are especially difficult to be met through a unified approach. Indeed, SKGs (which are a state-of-the-art approach to graph generation) are not even applicable to conditional edge prediction. Yet, FRFs are quite competitive in both link prediction and graph generation, even though from the statistical point of view they are much closer to ERGs than to the other models. These results are particularly encouraging, since they show how the nonparametric approach motivating the FRF model allows to accurately estimate network properties that are not aimed for explicitly in the model design. We interpret the superiority of FRFs with respect to other models from the exponential family as due on the one hand to the choice of a spectral approach in the design of the basic model statistic, and on the other hand to the nonparametric choice of modeling the involved potentials through MLP estimators. This suggests that focusing on such a low-level statistic as the Fiedler delta is a promising direction for building generative models capable of capturing a variety of network properties through a unified and compact approach.

7. Conclusions

The main motivation inspiring this work was the observation that current statistical models for large-scale networks make strong prior assumptions concerning the modeled network properties. Moreover, our experimental analysis showed how difficult it can be to accurately model at the same time the conditional independence structure underlying random networks and such high-level properties as the distributions of node degree and clustering coefficient. One key result emerging from the work we have presented is that the Laplacian spectrum of suitably sampled subgraphs is a very promising source of information for the purposes of statistical modeling. In particular, once exploited through the Fiedler delta statistic and the sort of Boltzmann distribution assumed in the FRF model, subgraph spectra allow to capture a variety of independence patterns and network properties through a more unified and more widely applicable approach than allowed by other models. In this regard, we stress the fact the the Fiedler delta statistic is just one possible way of extracting useful information from graph spectra. Clearly, many other network statistics could be defined, in principle, using the Laplacian spectrum. Since our contribution is nothing but a first attempt in this direction, we are encouraged to think that, if the goal is to uncover the low-level statistical laws underlying the structure of large-scale networks, then the Laplacian spectra of subgraph samples are a very promising place to look at.

Acknowledgments

This project has been supported by the French National Research Agency through program ANR-09-EMER-007. We are grateful to Gemma Garriga, Rémi Gilleron, Vanessa Sylvanie Kamga, Liva Ralaivola, Michal Valko, and Hyokun Yun for helping at different stages through general discussions, detailed suggestions, and specific remarks on various portions of this work. We are also indebted to the anonymous reviewers for their comments and criticisms on a preliminary draft of the paper.

References

Miguel Á. Carreira-Perpiñán and Geoffrey E. Hinton. On Contrastive Divergence Learning. In *Proceedings of the Tenth International Workshop on Articial Intelligence and Statistics (AISTATS 2005)*, pages 33–40, 2005.

Edoardo M. Airoldi, David M. Blei, Stephen E. Fienberg, and Eric P. Xing. Mixed Membership Stochastic Blockmodels. *Journal of Machine Learning Research*, 9:1981–2014, 2008.

Réka Albert and Albert-László Barabási. Statistical mechanics of complex networks. *Reviews of Modern Physics*, 74:47–97, 2002.

William N. Anderson and Thomas D. Morley. Eigenvalues of the Laplacian of a graph. *Linear and Multilinear Algebra*, 18:141–145, 1985.

Kendall E. Atkinson. *An Introduction to Numerical Analysis*. John Wiley & Sons, New York (NY), second edition, 1989.

Zhaojun Bai, James Demmel, Jack Dongarra, Axel Ruhe, and Henk van der Vorst, editors. *Templates for the Solution of Algebraic Eigenvalue Problems: A Practical Guide*. SIAM, Philadelphia (PA), 2000.

A.L. Barabási, H. Jeong, Z. Néda, E. Ravasz, A. Schubert, and T. Vicsek. Evolution of the social network of scientic collaborations. *Physica A*, 311:590–614, 2002.

Albert-László Barabási and Réka Albert. Emergence of scaling in random networks. *Science*, 286:509–512, 1999.

A. Barrat and M. Weigt. On the properties of small-world network models. *The European Physical Journal B*, 13:547–560, 2000.

Julian Besag. Spatial Interaction and the Statistical Analysis of Lattice Systems. *Journal of the Royal Statistical Society. Series B*, 36:192–236, 1974.

Aaron Clauset, Cosma Rohilla Shalizi, and Mark E.J. Newman. Power-Law Distributions in Empirical Data. *SIAM Review*, 51:661–703, 2009.

Dragoš M. Cvetković, Michael Doob, and Horst Sachs, editors. *Spectra of Graphs: Theory and Application*. Academic Press, New York (NY), 1979.

George Cybenko. Approximation by Superpositions of a Sigmoidal Function. *Mathematics of Control, Signals, and Systems*, 2:303–314, 1989.

V.A. Epanechnikov. Nonparametric Estimation of a Multidimensional Probability Density. *Theory of Probability and its Applications*, 14:153–158, 1969.

P. Erdős and A. Rényi. On Random Graphs, I. *Publicationes Mathematicae Debrecen*, 6: 290–297, 1959.

Miroslav Fiedler. Algebraic connectivity of graphs. *Czechoslovak Mathematical Journal*, 23: 298–305, 1973.

Ove Frank and David Strauss. Markov Graphs. *Journal of the American Statistical Association*, 81:832–842, 1986.

Antonino Freno, Mikaela Keller, Gemma C. Garriga, and Marc Tommasi. Spectral Estimation of Conditional Random Graph Models for Large-Scale Network Data. In *Proceedings of the 28th Conference on Uncertainty in Artificial Intelligence (UAI 2012)*, pages 265–274. AUAI Press, 2012.

Anna Goldenberg, Alice X. Zheng, Stephen E. Fienberg, and Edoardo M. Airoldi. A Survey of Statistical Network Models. *Foundations and Trends in Machine Learning*, 2(2):129–233, 2009.

Geoffrey E. Hinton. Training Products of Experts by Minimizing Contrastive Divergence. *Neural Computation*, 14(8):1771–1800, 2002.

Peter D. Hoff, Adrian E. Raftery, and Mark S. Handcock. Latent Space Approaches to Social Network Analysis. *Journal of the American Statistical Association*, 97(460):1090–1098, 2002.

Kurt Hornik. Approximation capabilities of multilayer feedforward networks. *Neural Networks*, 4(2):251–257, 1991.

Christian Hübler, Hans-Peter Kriegel, Karsten M. Borgwardt, and Zoubin Ghahramani. Metropolis Algorithms for Representative Subgraph Sampling. In *Proceedings of the 8th IEEE International Conference on Data Mining (ICDM)*, pages 283–292. IEEE Computer Society, 2008.

Eric D. Kolaczyk. *Statistical Analysis of Network Data. Methods and Models*. Springer, New York (NY), 2009.

Jure Leskovec and Christos Faloutsos. Sampling from large graphs. In *Proceedings of the Twelfth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2006)*, pages 631–636, 2006.

Jure Leskovec, Jon M. Kleinberg, and Christos Faloutsos. Graphs over Time: Densification Laws, Shrinking Diameters and Possible Explanations. 2005.

Jure Leskovec, Jon M. Kleinberg, and Christos Faloutsos. Graph evolution: Densification and shrinking diameters. *Transactions on Knowledge Discovery from Data*, 1(1), 2007.

Jure Leskovec, Deepayan Chakrabarti, Jon Kleinberg, Christos Faloutsos, and Zoubin Ghahramani. Kronecker graphs: An approach to modeling networks. *Journal of Machine Learning Research*, 11:985–1042, 2010.

Bojan Mohar. The Laplacian Spectrum of Graphs. In Y. Alavi, G. Chartrand, O. R. Oellermann, and A. J. Schwenk, editors, *Graph Theory, Combinatorics, and Applications*, pages 871–898. Wiley, 1991.

Mark E.J. Newman. Clustering and preferential attachment in growing networks. *Physical Review E*, 64:025102, 2001.

Mark E.J. Newman. *Networks. An Introduction*. Oxford University Press, New York (NY), 2010.

Andrew Y. Ng, Michael I. Jordan, and Yair Weiss. On Spectral Clustering: Analysis and an algorithm. In *Advances in Neural Information Processing Systems (NIPS)*, pages 849–856. MIT Press, 2001.

Emanuel Parzen. On Estimation of a Probability Density Function and Mode. *Annals of Mathematical Statistics*, 33:1065–1076, 1962.

Garry L. Robins, Tom Snijders, Peng Wang, Mark S. Handcock, and Philippa E. Pattison. Recent developments in exponential random graph (p^*) models for social networks. *Social Networks*, 29:192–215, 2007.

Murray Rosenblatt. Remarks on Some Nonparametric Estimates of a Density Function. *The Annals of Mathematical Statistics*, 27(3):832–837, 1956.

Tom A. B. Snijders, Philippa E. Pattison, Garry L. Robins, and Mark S. Handcock. New Specifications for Exponential Random Graph Models. *Sociological Methodology*, 36:99–153, 2006.

Ulrike von Luxburg. A tutorial on spectral clustering. *Statistics and Computing*, 17:395–416, 2007.

Duncan J. Watts and Steven H. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393:440–442, 1998.

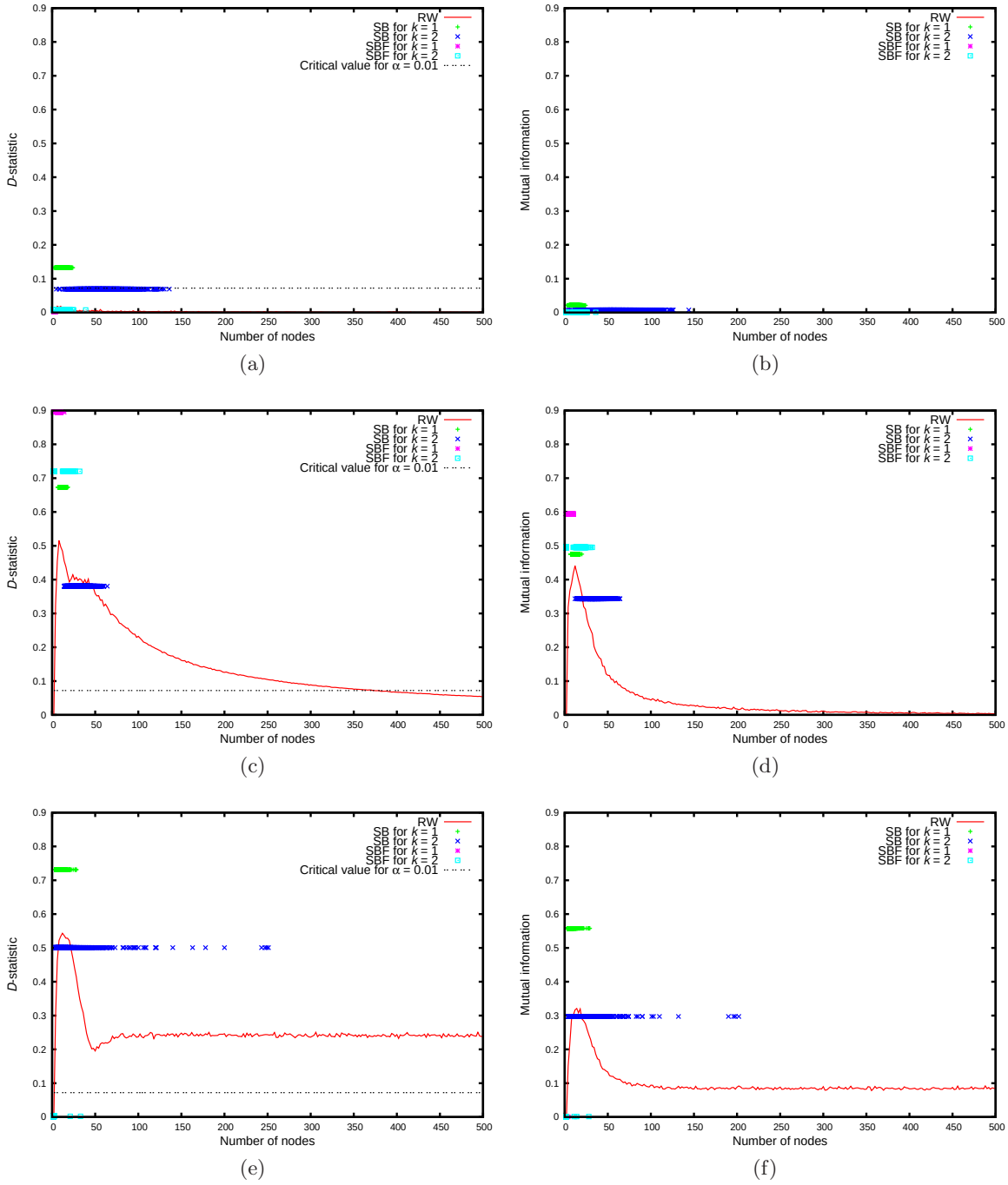


Figure 2: Kolmogorov-Smirnov D-statistic and mutual information values (measured for p_0 vs. p_1 and $\Delta\lambda_2(u, v, \mathcal{G}_{uv})$ vs. X_{uv} respectively) for the Synth-ER (a)–(b), Synth-WS (c)–(d), and Synth-BA (e)–(f) networks. Snowball samples are plotted as individual points because we do not have direct control of subgraph size (which is instead the case for random walk samples).

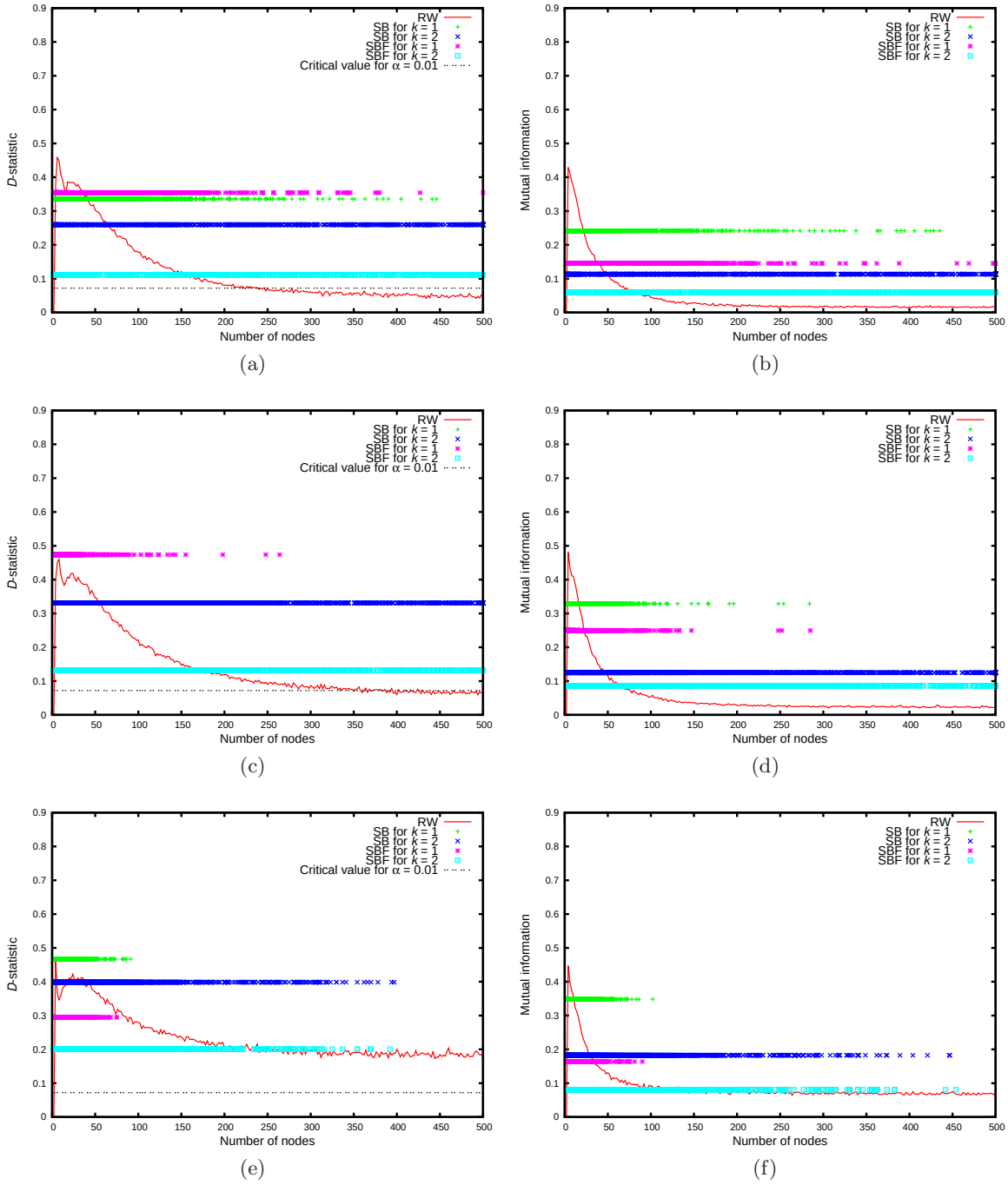


Figure 3: Kolmogorov-Smirnov D-statistic and mutual information values (measured for p_0 vs. p_1 and $\Delta\lambda_2(u, v, \mathcal{G}_{uv})$ vs. X_{uv} respectively) for the CA-AstroPh (a)–(b), CA-CondMat (c)–(d), and CA-GrQc (e)–(f) networks. Snowball samples are plotted as individual points because we do not have direct control of subgraph size (which is instead the case for random walk samples).

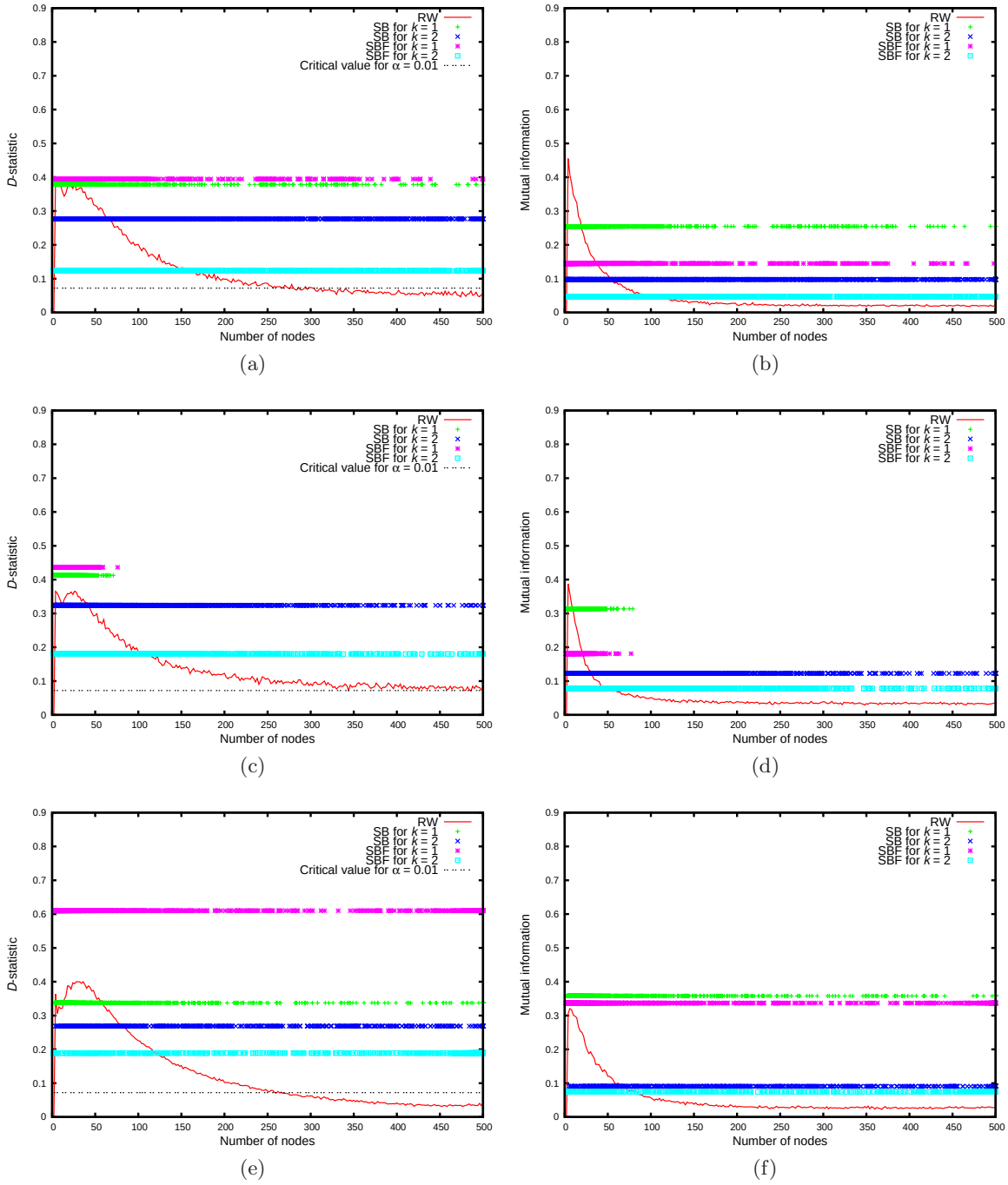


Figure 4: Kolmogorov-Smirnov D-statistic and mutual information values (measured for p_0 vs. p_1 and $\Delta\lambda_2(u, v, \mathcal{G}_{uv})$ vs. X_{uv} respectively) for the CA-HepPh (a)–(b), CA-HepTh (c)–(d), and AS-Skitter (e)–(f) networks. Snowball samples are plotted as individual points because we do not have direct control of subgraph size (which is instead the case for random walk samples).

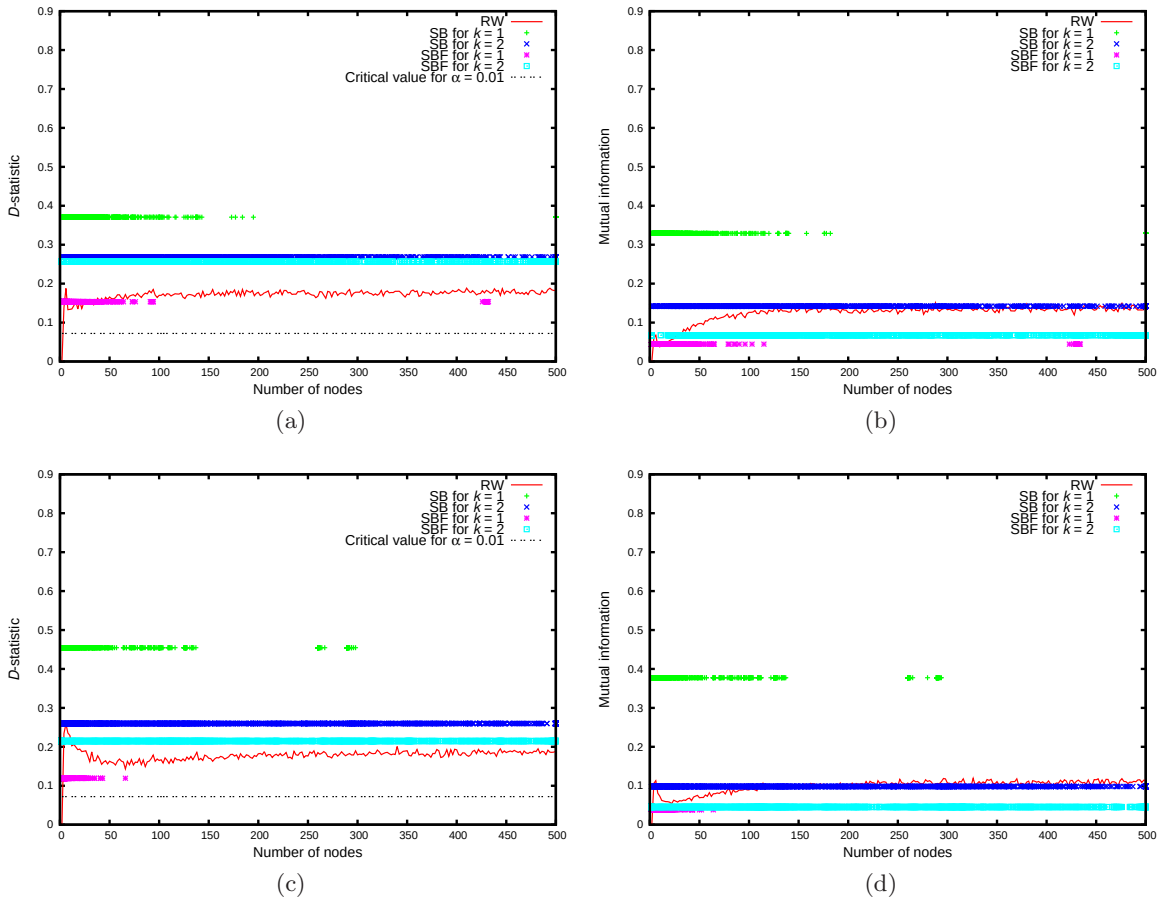


Figure 5: Kolmogorov-Smirnov D-statistic and mutual information values (measured for p_0 vs. p_1 and $\Delta\lambda_2(u, v, \mathcal{G}_{uv})$ vs. X_{uv} respectively) for the PPI-Drosoph (a)–(b) and PPI-Sacchar (c)–(d) networks. Snowball samples are plotted as individual points because we do not have direct control of subgraph size (which is instead the case for random walk samples).

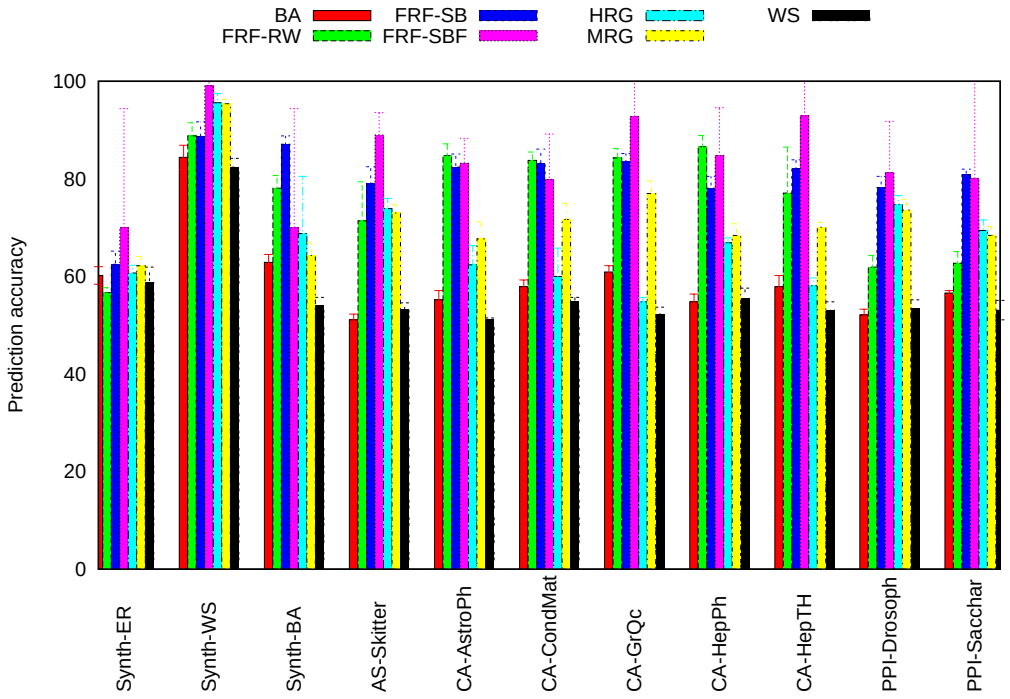


Figure 6: Average link prediction accuracy (%) measured by 5-fold cross-validation for the networks listed in Table 1.

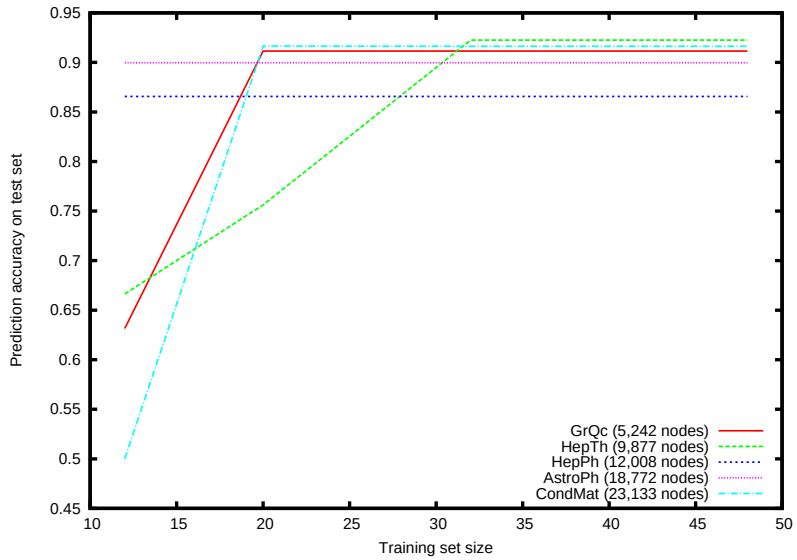
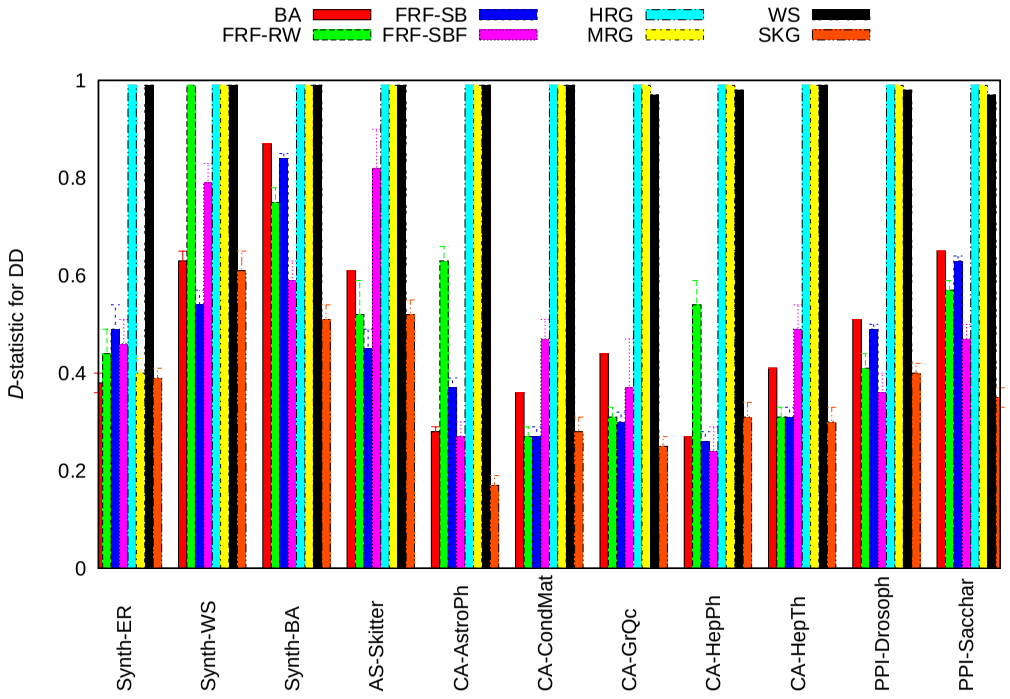
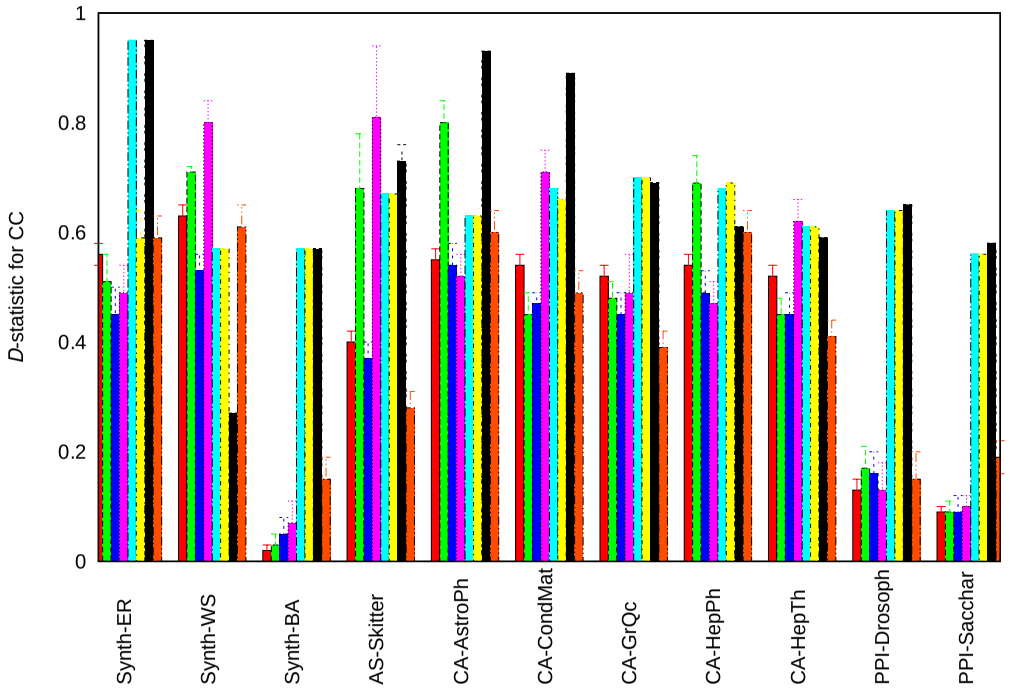


Figure 7: Prediction accuracy of FRFs (using one-wave snowball sampling) on the arXiv networks for a growing training set size.



(a)



(b)

Figure 8: Average D-statistic values for degree distribution and clustering coefficient on the networks listed in Table 1.